Learning from Data Heterogeneity: Algorithms and Applications

Jingrui He
School of Computing, Informatics, and Decision Systems Engineering, Arizona State University
jingrui.he@asu.edu

Abstract

Nowadays, as an intrinsic property of big data, *data heterogeneity* can be seen in a variety of real-world applications, ranging from security to manufacturing, from healthcare to crowdsourcing. It refers to any inhomogeneity in the data, and can be present in a variety of forms, corresponding to different types of data heterogeneity, such as task/view/instance/oracle heterogeneity. As shown in previous work as well as our own work, learning from data heterogeneity not only helps people gain a better understanding of the large volume of data, but also provides a means to leverage such data for effective predictive modeling. In this paper, along with multiple real applications, we will briefly review state-of-the-art techniques for learning from data heterogeneity, and demonstrate their performance at addressing these real world problems.

1 Introduction

A common, fundamental property of many data mining applications is heterogeneity, which refers to any inhomogeneity in the data. For example, the target application may consist of multiple heterogeneous data sets with varying data distributions (task heterogeneity); each example may be characterized using features from heterogeneous sources (view heterogeneity); an example might be further decomposed into multiple components (instances) with heterogeneous labels (instance heterogeneity); the data labels may be provided by multiple heterogeneous oracles (oracle heterogeneity). To model data heterogeneity, various techniques have been proposed in the past decades, mainly focusing on a single type of heterogeneioty. For example, multi-task learning [Caruana, 1997] and transfer learning [Pan and Yang, 2010] aims to model task heterogeneity; multi-view learning aims to model view heterogeneity [Xu *et al.*, 2013]; multi-instance learning aims to model instane heterogeneity [Zhou, 2004]; and crowd sourcing aims to model oracle heterogeneity [Gao *et al.*, 2016; Shah *et al.*, 2015].

Furthermore, many emerging high impact applications bring a new challenge, i.e., different types of heterogeneity often co-exist in these applications. For example, for abnormal user detection in the financial world, each user is characterized by diverse types of information (view heterogeneity), such as the demographic and financial information; he or she might have a set of accounts across different platforms with varying levels of suspicion (instance heterogeneity); the patterns of different abnormal user groups (e.g., ID theft vs. synthetic ID) might be correlated with each other (task heterogeneity). Another exampe is quality control in manufacturing processes, where data heterogeneity is reflected in the hierarchical structure made up by multiple tools and chambers bearing similar configurations (task heterogeneity), the multi-step nature of the manufacturing process with diverse impact on the quality (instance heterogeneity), the different kinds of process variables recorded in each step (view heterogeneity), etc. For such applications, we aim to explore the interplay among multiple types of data heterogeneity in such a way that outperforms modeling each type of heterogeneity separately.

For the rest of this paper, we will demonstrate the different types of data heterogeneity in multiple application domains, and introduce state-of-the-art techniques for modeling data heterogeneity that have been successfully applied in these domains.

2 Applications and Algorithms

In this section, I will instantiate various types of data heterogeneity associated with different real applications, including abnormal user detection, manufacturing, and healthcare, briefly review state-of-the-art techniques for learning from such heterogeneity or combination of multiple types of heterogeneity, and provide empirical evidence showing the effectivenss of these techniques.

2.1 Abnormal User Detection

View Heterogeneity

In abnormal user detection, it is often the case that each user is associated with multiple types of information, corresponding to multiple views. For example, in the financial world, each user would have information from numerous financial transactions, customer profiles, social media, etc.

To model the view heterogeneity in abnormal user deteciton, in [Zhou *et al.*, 2015], we proposed a novel framework named *MUVIR* for detecting the initial examples from

the rare classes (corresponding to abnormal users) in the presence of multi-view data. The key idea is to integrate view-specific posterior probabilities of the example coming from the rare class given features from each view, in order to obtain the estimate of the overall posterior probability given features from all the views. *MUVIR* is essentially a wrapper in the sense that the view-specific posterior probabilities can be inferred from the scores computed using a variety of existing techniques, such as [He, 2012]. Furthermore, it can be generalized to handle problems where the exact priors of the rare classes (as required by existing techniques for computing the score from each individual view), or the proportions of abnormal users, are unknown.

In Figure 1, we compared *MUVIR* and its variant *MUVIR-LI* that does require the exact priors of the rare classes with two existing rare category detection methods [He *et al.*, 2008] designed for a single view on the Statlog data set. As we can see, *MUVIR* is able to identify the initial examples from all the classes (especially the rare ones) with less label requests as compared with GRADE [He *et al.*, 2008], although the information provided to both algorithms is exactly the same. Similarly, *MUVIR-LI* also outperforms GRADE-LI [He *et al.*, 2008], both requiring inexact priors of the rare classes. These results show that leveraging the view heterogeneity can improve the performance of abnormal user detection, especially the detection of the initial examples of a new abnormal class.

Label Heterogeneity

For detecting abnormal patterns in temporal data, we can often observe a bi-level structure as shown in Figure 2: of the large number of temporal sequences, only a few of them are abnormal; within the abnormal temporal sequences, the abnormal patterns may only be present in a few time segments and are similar among themselves, forming a rare category of temporal patterns. Tailored for such bi-level structures, we proposed the *BIRAD* algorithm for detecting the abnormal temporal patterns [Zhou *et al.*, 2016].

The key idea of *BIRAD* [Zhou *et al.*, 2016] is to maximize the likelihood of observing the data on both the sequence-level and the segment-level. Furthermore, it uses sequence-specific hidden Markov models to generate segment-level labels, and leverages the similarity among the abnormal time segments to estimate the model parameters. To solve the optimization problem, *BIRAD* repeatedly updates the sequence-level labels, segment-level labels, and the model parameters via Block Coordinate Descent until convergence.

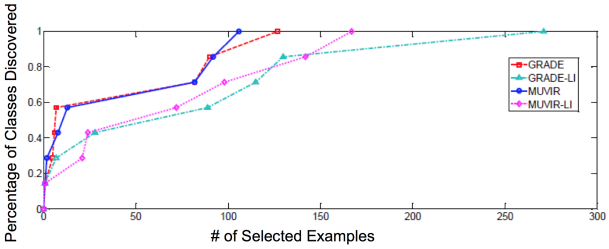

Figure 1: Comparison on Statlog Data Set.

View and Label Dual Heterogeneity

To jointly model the view and label dual heterogeneity in abnormal user detection, in [Zhang *et al.*, 2013], we proposed Multi-Instance Learning from Multiple Information Sources, or MI^2LS, as well as its speed-up version FMI^2LS. It combines Constraint Concave-Convex Programming (CCCP) method and an adapted Stoachastic Gradient Descent (SGD) method, which demonstrated promising results in insider threat detection, as shown in Figure 3.

2.2 Manufacturing

Task (Structure) Heterogeneity

In semiconductor manufacturing, to produce a certain IC (Integrated Circuit) device, multiple tools will be deployed following the same recipe process, and each tool has multiple chambers to carry out the task. During the process, various process variables with be monitored and recorded over time, producing huge amount of time series data. The time series data naturally fit into a two-dimensional array, or data matrix, where each row of the array corresponds to one chamber, each column corresponds to one process variable, and each element in this array corresponds to the measurements of the process variable over time. Such structural temporal data contain rich information about the manufacturing process, and thus can be exploited to help domain experts gain more insights into the recipe of the IC device.

In particular, if we simultaneously cluster the rows and colums of the data matrix in such a way that groups similar chambers and process varaibles together, we would be able to uncover the heterogeneous structure of the data matrix. In other words, we would be able to identify chambers with similar behaviors and process variables with similar patterns over time; we would also be able to identify the problematic chambers and process variables for the sake of fault detection.

To this end, we proposed the *C-Struts* Framework [Zhu and He, 2016] to model the structure heterogeneity. In this framework, we first interpreted the structural information associated with the two-dimensional array as a set of constraints on the cluster membership. Then we introduced an auxiliary probability distribution taking these constraints into consideration, analyzed its properties, and built a prototype for each row/ column accordingly. Finally, we ued an iterative procedure to repeatedly assign each row/column to the closest prototype. *C-Struts* has been extensively tested on benchmark data sets. In particular, Figure 4 shows the comparison results on a semiconductor data set.

Label (Output) Heterogeneity

Another important application of co-clustering is to model the label (output) heterogeneity in cargo pricing optimization [Zhu *et al.*, 2015], where the goal is to predict both the optimal price for the bid stage and the outcome of the transaction (win rate) in the decision stage with respect to each origination/ destination pair. To this end, we proposed a probabilistic framework to simul- taneously construct dual predictive models and uncover the co-clusters of originations and destinations. It maximizes the conditional probability of observing the responses from both the quotation stage and the decision stage, taking into consideration both the features and

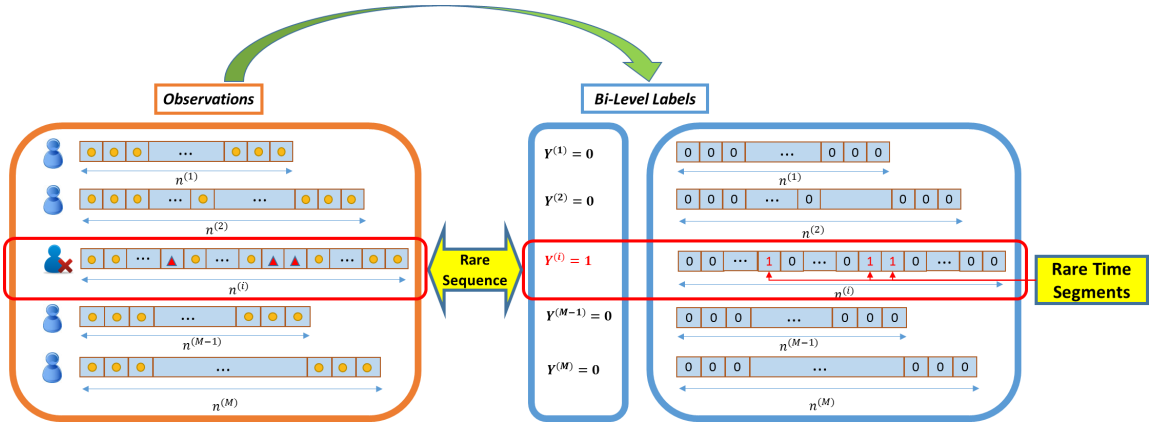

Figure 2: Illustration of the Bi-Level Structure.

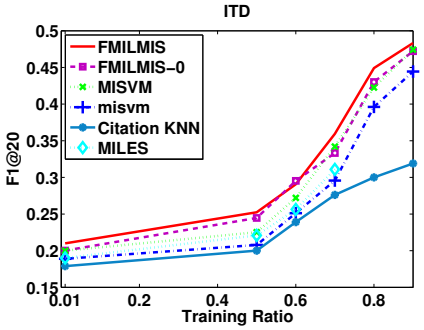

Figure 3: Comparison results on insider threat detection data.

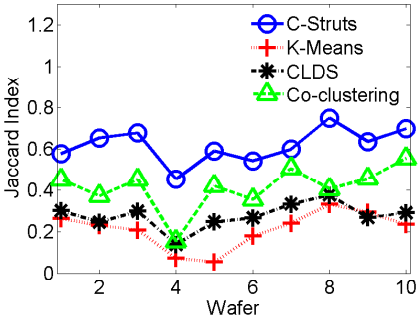

Figure 4: Comparison results on semiconductor data set.

the co-clusters of the origination/ destination pair. To minimize the objective function, we proposed the *COCOA* algorithm, which generates both the suite of predictive models for all the pairs of originations and destinations, as well as the co-clusters consisting of similar pairs. Figure 5 shows that the proposed *COCOA* algorithm outperforms HGLM [Albert, 1988] in terms of the revenue.

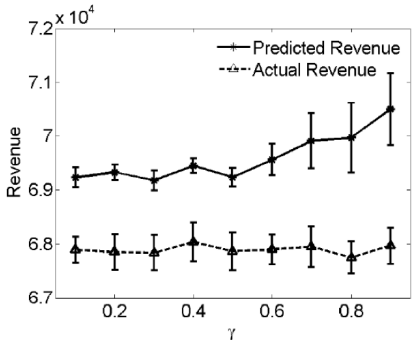

Figure 5: Comparison of revenue for action prediction in computational advertisement.

2.3 Healthcare

Task Heterogeneity

Nowadays, for many chronic diseases such as diabetes mellitus, numerous disease-specific social networks have been created to faciliate information and resource sharing among the patients. In many cases, finding other patients with similar conditions, symptoms, questions and concerns can enable the formation of patient support groups, which in turn helps the patients manage their conditions over time. On the other hand, such disease-specific social networks are often isolated from one another, which creates a virtual barrier for patient communication across multiple social networks.

To address this problem, in [Nelakurthi and He, 2017], motivated by the fact that social networks dedicated to the same disease tend to share the same topics as well as the interests of users groups in certain topics, we proposed to jointly decompose the user-keyword matrices from multiple social networks in such a way that the topics and user group-topic association matrices are shared. Finally, based on the learned user represention in the latent feature space, we make use of random walk with restart to estimate the similarity among users across different social networks, and make recommendations accordingly. Figure 6 illustrates the key idea of the proposed

technique.

View and Label Dual Heterogeneity

For diabetes patients with multiple biomarkers measured over time, an important question is how to determine combination of treatments based on these biomarker measurements. In particular, this problem exhibits both view and label dual heterogeneity: the various types of biomarker measurements correspond to multiple views, and each type of treatment correspond to one label. Therefore, our goal here is to effectively leverage both types of heterogeneity to build reliable models.

To address this problem, in [Yang *et al.*, 2016; Yang and He, 2015; Yang *et al.*, 2014], we proposed a hypergraph-based framework to model 3 types of correlations, namely example-to-example, label-to-label, and view-to-view correlations. Then we formualted an optimization problem based on this hypergraph, where the objective function consists of 4 terms: example consistency on the graph, label correlation, view consistency, as well as the empirical loss. In particular, the label correlation term measures the similarity of the label-specific classifiers per view, whereas the view consistency term compares the output from each view with respect to all the labels. Figure 7 shows the comparison results between the propsed framework named L^2F and existing work on a diabetes data set, where L^2F outperforms the other methods in terms of the F-score.

3 Conclusion

In this paper, we focused on one important aspect of big data, i.e., data heterogeneity. It is present in many high impact applications in a variety of forms, corresponding to different types of heterogeneity. Although data heterogneity has been studied in the past decades, new challenges arising from these applications call for new algorithms and theories, such as the effective joint modeling of multiple types of heterogeneity, the relationship between model complexity and model performance, etc. We believe that advances in these aspects will not only benefit artificial intelligence in general, but also provide more powerful tools for a variety of application domains such as security, manufacturing, healthcare, etc.

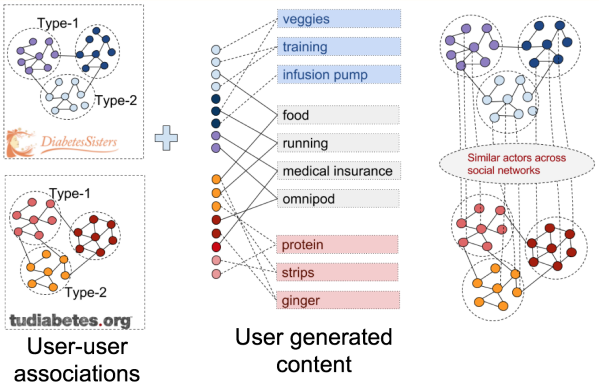

Figure 6: Cross network link recommendation.

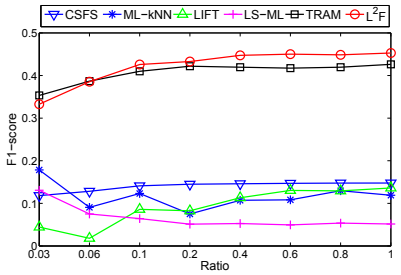

Figure 7: Comparison of the F-score on diabetes data.

Acknowledgments

This work is supported by National Science Foundation under Grant No. IIS-1552654, ONR under Grant No. N00014-15-1-2821, and an IBM Faculty Award. The views and conclusions are those of the authors and should not be interpreted as representing the official policies of the funding agencies or the government.

References

[Albert, 1988] James H. Albert. Computational methods using a bayesian hierarchical generalized linear model. *Journal of the American Statistical Association*, 83(404):1037–1044, 1988.

[Caruana, 1997] Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

[Gao *et al.*, 2016] Chao Gao, Yu Lu, and Dengyong Zhou. Exact exponent in optimal rates for crowdsourcing. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 603–611, 2016.

[He *et al.*, 2008] Jingrui He, Yan Liu, and Richard D. Lawrence. Graph-based rare category detection. In *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008), December 15-19, 2008, Pisa, Italy*, pages 833–838, 2008.

[He, 2012] Jingrui He. *Analysis of Rare Categories*. Cognitive Technologies. Springer, 2012.

[Nelakurthi and He, 2017] Arun Reddy Nelakurthi and Jingrui He. Finding cut from the same cloth: Cross network link recommendation via joint matrix factorization. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, pages 1467–1473, 2017.

[Pan and Yang, 2010] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.*, 22(10):1345–1359, 2010.

[Shah *et al.*, 2015] Nihar B. Shah, Dengyong Zhou, and Yuval Peres. Approval voting and incentives in crowdsourcing. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 10–19, 2015.

[Xu *et al.*, 2013] Chang Xu, Dacheng Tao, and Chao Xu. A survey on multi-view learning. *CoRR*, abs/1304.5634, 2013.

[Yang and He, 2015] Pei Yang and Jingrui He. A graph-based hybrid framework for modeling complex heterogeneity. In *2015 IEEE International Conference on Data Mining, ICDM 2015, Atlantic City, NJ, USA, November 14-17, 2015*, pages 1081–1086, 2015.

[Yang *et al.*, 2014] Pei Yang, Jingrui He, Hongxia Yang, and Haoda Fu. Learning from label and feature heterogeneity. In *2014 IEEE International Conference on Data Mining, ICDM 2014, Shenzhen, China, December 14-17, 2014*, pages 1079–1084, 2014.

[Yang *et al.*, 2016] Pei Yang, Hongxia Yang, Haoda Fu, Dawei Zhou, Jieping Ye, Theodoros Lappas, and Jingrui He. Jointly modeling label and feature heterogeneity in medical informatics. *TKDD*, 10(4):39:1–39:25, 2016.

[Zhang *et al.*, 2013] Dan Zhang, Jingrui He, and Richard D. Lawrence. MI2LS: multi-instance learning from multiple informationsources. In *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013, Chicago, IL, USA, August 11-14, 2013*, pages 149–157, 2013.

[Zhou *et al.*, 2015] Dawei Zhou, Jingrui He, K. Seluk Candan, and Hasan Davulcu. MUVIR: multi-view rare category detection. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 4098–4104, 2015.

[Zhou *et al.*, 2016] Dawei Zhou, Jingrui He, Yu Cao, and Jae-sun Seo. Bi-level rare temporal pattern detection. In *IEEE 16th International Conference on Data Mining, ICDM 2016, December 12-15, 2016, Barcelona, Spain*, pages 719–728, 2016.

[Zhou, 2004] Zhi-Hua Zhou. Multi-instance learning: A survey. Computer sciences technical report, Nanjing University, 2004.

[Zhu and He, 2016] Yada Zhu and Jingrui He. Co-clustering structural temporal data with applications to semiconductor manufacturing. *TKDD*, 10(4):43:1–43:18, 2016.

[Zhu *et al.*, 2015] Yada Zhu, Hongxia Yang, and Jingrui He. Co-clustering based dual prediction for cargo pricing optimization. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*, pages 1583–1592, 2015.